基础

> 简介：多领域交叉，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论。
> 研究计算机怎样模拟或者实现人类学习行为，获取新知识或者技能，重新组织已有的知识结构使之不断改善自身性能。

一、简介、概念

> 广义：赋予机器以学习的能力，使机器能够完成通过编程无法直接完成的功能的方法。
> 具体到实践: 通过利用数据，训练出模型，然后使用模型预测的方法。

也就是在给定的数据或者样本的条件下，经过模型训练，进行预测或者决策的过程。
从这个角度来说，机器学习就是计算机模拟人类学习行为的一个过程，从数据学习中得出规律或者模式，并把规律和模式应用在新数据上完成预测的任务。

复杂性：在于需要根据大量已有样本或者数据集，从中找出规律或者模式，从而得到最优参数以应用在新数据上进行预测。可能需要经过成千上万次的计算，可能需要采用各种类型的算法，包含大量概率论、统计学、凸分析等知识。

本质：通过机器的计算来模拟人类思考的过程，不同之处在于机器学习需要的输入、训练过程、对未知的预测都是定量化、程序化、模式化的，这要是局限。

与人类思维相比：人类的思考有感性的东西，在面对非结构性、不规律的问题和场景时，能够找到规律并作出判断，但是信息量较大的时候，难以找到规律，这就需要机器学习帮助。

1.1 机器学习思维

机器学习最核心、最重要的环节-> 模型训练的方式和方法。

一个著名的定理: 针对某一领域的所有问题，所有算法的期望性能是相同的。

> 也就是在某一个领域，如果就一个问题算法A比算法B更优，那么一定存在别的问题，算法B比算法A更优。这个定理说明，没有一个可靠的算法吃遍天下的可能性，都需要具体问题具体分析。

1.2 几种常见的算法

1.2.1 K近邻算法

> 近朱者赤近墨者黑。

先根据已知样本确定每一类别的划分区域，再根据未知类别样本与哪个类别的划分区域最近，就将该样本归于哪类。

1.2.2 决策树

根据每一个决策问题的自然状态或者条件出现的概率、益损值、预测结果，来做出不同的决策，再从这一决策出发遇到该决策条件下不同选择问题的时候，又进一步根据其概率、益损值、预测结果等进一步划分，由此形成不同问题的分类。

1.2.3 朴素贝叶斯

> 吃一堑长一智。

通过后验经验法（条件概率）来对未知进行预测。

1.2.4 线性回归

将不同的变量与因变量建立一条线性映射，接下来就是通过训练，找到一条预测偏差最小的直线。

1.2.5 SVN支持向量机

他把数据映射到多维空间中以点的形式存在，然后找到能够分类的最优超平面，最后根据这个平面分类。

1.2.6 K均值聚类

> 物理类聚，人以群分

与K近邻算法相似，根据距离的思想定义类别标准，只不过它没有预先分类，需要自己去确定，后者是根据已有类别划分，建立规则去预测未知类别。

二、基本框架体系

2.1 主流分类

- 有监督。
  > 在已知变量标识值的引导下，来预测目标变量的值或所属类型，并通过不断调整模型的参数已达到更高的准确率。
  > 在监督学习中，输入的数据被称为"训练数据"，每组训练数据有一个明确的标识或结果。
  > 在建立预测模型的时候，有监督学习建立一个学习过程，将预测结果与"训练数据"的实际结果进行比较，不断调整预测模型，直到模型的预测结果达到一个预期的准确率。
  > 应用场景：分类和回归问题。

- 无监督。

  > 在缺乏目标变量标识值的条件下，对数据进行训练和学习，来推导出概括数据潜在联系的模式，如寻找数据中的相关关系，描述数据的趋势，对数据中的不同簇进行聚类，寻找数据中的异常值。
  > 数据不背特别标识，学习模型是为了推断出数据的一些内在结构。它的目标不是告诉计算机怎么做，而是让他自己学习如何做。
  > 应用场景：关联规则的学习，聚类。

区别在于输入的数据是否被标记。

2.2 框架体系分类

> 分类、回归、聚类分析、关联规则。

2.2.1 分类

分类任务是确定对象术语哪个预定义的目标类，属于有监督学习。